DOCUMENT RESUME

ED 059 749 LI 003 488

TITLE	Guidelines for the Establishment and Development of Monolingual Thesauri for Information Retrieval.
INSTITUTION	United Nations Educational, Scientific, and Cultural Organization, Paris (France).
REPORT NO	SC-WS-500
PUB DATE	22 Dec 71
NOTE	30p.; (36 References)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Development; Guidelines; *Information Retrieval; Information Storage; Lexicography; *Subject Index Terms; *Thesauri

ABSTRACT

These guidelines are applicable in the sector of information and documentation for the preparation and development of monolingual thesauri for information storage and retrieval, irrespective of the technical field being dealt with. No provision has been made for mathematical and structural chemical formulae. The guidelines are drafted for general application, and are intended to facilitate the preparation and development of thesauri regardless of whether they are administered mechanically or manually. (Author/SJ)

SC/WS/500
PARIS, 22 December 1971
Original: English

UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

Guidelines for the Establishment and Development of Monolingual Thesauri for Information Retrieval

(This document is the revised version of document UNESCO/SC/MD 20, Paris, 6 July 1970, Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval).

HISTORICAL NOTE

These Guidelines are the result of a joint effort of Unesco and ISO. In view of the important rôle which thesauri play as a tool in the information process, both Organizations recognized the need for co-operation in order to avoid duplication of work.

During the ISO/TC 46 meeting in Lisbon 1971, it was therefore resolved that ISO and Unesco should jointly draft a document entitled "Guidelines for the preparation of Monolingual Thesauri", which would serve on one side as an ISO draft proposal and, on the other, as a revised version of the Unesco Guidelines.

A joint Unesco-ISO/TC 46 Task Group prepared the present "Guidelines for the Establishment and Development of Monolingual Thesauri for Information Retrieval" which are mainly based on the following documents:

Unesco, Guidelines for the establishment and development of monolingual scientific and technical thesauri for information retrieval. Paris, 6 July 1970.

The German draft proposal, Thesaurus Guidelines, Dec. 1970, ISO/TC 46 (Germany-2) 984E.

The first draft of the American National Standard Guidelines for Thesaurus Structure, Construction and Use. ANSI Z39, Feb.1971.

The comments of:

ISO/INFCO, (Observations of ISO drafted by the secretariat of ISO/TC 46, Comments proposed by INFCO/WG 1 "Indexing of Standards" to "Unesco guidelines..." ISO/INFCO (TC 46-2) 62E); Denmark (Remarks to ISO/INFCO (WG 1-2) 54 and to some other parts of the Unesco Guidelines, 3rd Draft (Feb.1970). Annex to Doc. ISO/INFCO (TC 46-2) 62E; France (Commentaires particuliers sur le document 984E "Thesaurus Guidelines" pris comme base de travail, ISO/TC 46 (France 3) 1003F); Romania (Comments on the Draft Proposal for "Thesaurus Guidelines" ISO/TC 46 (Germany-2) 984E ISO/TC 46 (Romania-2), 1015E; USSR (Comments proposed by the USSR member body to the documents: ISO/INFCO (TC 46-2) 62E and ISO/TC 46 (Germany-2) 984E), were taken into account as well as the proceedings of the International Conference on General Principles of Thesauri Building, Warsaw, March 1970. Furthermore, the conclusions of the ISO/TC 46 Ad Hoc Working Group on Thesaurus Construction in Lisbon 1971 were incorporated as well as the comments of many experts and competent organizations.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 SPHERE OF APPLICATION

These guidelines are applicable in the sector of information and documentation for the preparation and development of monolingual thesauri for information storage and retrieval, irrespective of the technical field being dealt with. No provision has been made for mathematical and structural chemical formulae. The guidelines are drafted for general application. They should be completed in accordance with the characteristics of each particular language.

1.2 AIM

In information and documentation centres there is a need for practical methods of representing concepts simply and clearly and of ordering them by clarifying their interrelationships. By applying these methods and the terminological control of concepts, it is possible to introduce an effective and cost-saving system of analysis, classification and retrieval in documentation systems.

Agencies in many countries have either built or are about to build thesauri for their own purposes. It is obvious that the transfer of information requires a standard set of rules for the construction of thesauri. These guidelines will be useful:

for a single agency, in order to facilitate construction and development of a thesaurus;

to facilitate the transfer of information between agencies.

The purpose of these guidelines is to facilitate the preparation and development of thesauri regardless of whether they are administered mechanically or manually. The guidelines were, therefore, prepared as an attempt to lay the basis for compatibility, both at present and in the future, of thesauri that are being elaborated simultaneously in most of the disciplines of science, basic as well as applied.

2. DEFINITIONS

2.1 DEFINITION OF THESAURUS

A thesaurus may be defined either in terms of its function or its structure. In terms of function, a thesaurus is a terminological control device used to translate from the natural language of documents, indexers, or users into a more constrained "system language" (documentation language, information language), as well as to translate the system language back into natural language.

In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which comprehensively covers a specific domain of knowledge.

As a structured subset of natural language it describes the subject content of documents, works of art, or collections of data. The language needed to describe the document or object - non-subject terms such as names of authors, corporate authors, journals, museums, locations, etc. (bibliographic description) - need not be a part of a formalized thesaurus though such terms are usually necessary for identification or location of that which is indexed. They may, in fact, comprise an adjunct to a thesaurus.

A particular thesaurus should accurately reflect the information content of the body of documents or other items in a collection to which the thesaurus applies. It should contain terms and cross-references appropriate to the subject matter, taking into consideration both the language of the document collection and the language and the information needs of the users.

Based on the nature of the terminological control, two major types of thesauri are currently being developed:

(a) thesauri which perform terminological control by preferred terms, i.e. thesauri in which only one of the terms denoting a concept is permitted for indexing and retrieval, and

(b) thesauri which perform terminological control by allowing all terms denoting a concept to be used for indexing and retrieval, but which assign these terms to another unambiguous representation of the concept (e.g. concept number, notation).(*) This type of thesaurus makes it easier to cancel earlier synonyms and concept relationship if a term changes its meaning.

2.2 RELATIONS TO OTHER DOCUMENTATION LANGUAGES

In contrast to a dictionary, which provides definitions for given words or terms, a thesaurus provides words or terms to express meanings that are implied by the term relationships given in the thesaurus.

A thesaurus may be arranged like an alphabetical index and the terms in a thesaurus may be used to construct an index. But the thesaurus itself is not an index. An index to a collection must have addresses or locators for items in the collection associated with each term, but a thesaurus contains the terms only, without the addresses or locators of an index. A thesaurus classifies terms by arranging them in hierarchical classes. As a "term classification system" a thesaurus has some similarities with subject matter classification systems, as represented by the Universal Decimal Classification. But whereas hierarchical subject classification systems try to show the whole system of hierarchical

(*) While thesauri using preferred terms can be maintained manually, thesauri not using preferred terms require machine maintenance and retrieval.

relations, the thesaurus shows relations necessary for indexing and retrieval according to the body of documents and the information needs of users.

A thesaurus is one kind of authority list, that is, the preferred terms in a particular thesaurus are required indexing and retrieval terms for a given information and documentation system. There are other kinds of natural-language-based authority lists, such as subject heading lists. In general, however, these do not have the hierarchical structure of the thesaurus.

3. THESAURUS STRUCTURE

3.1 DESCRIPTORS

The structure of a thesaurus is the internal form of individual entries and the arrangement of the various entries in relation to one another. Cross-references in a thesaurus make explicit the ways in which entries relate to each other in a network of concepts.

The terms permitted by a thesaurus for use in indexing are called descriptors. In thesauri using preferred terms, the descriptors are the preferred terms. The descriptor can be characterized as an authorized and formulized term or symbol in a thesaurus, used to represent unambiguously the concepts of documents and queries.

With regard to the aim of international co-operation, two levels of descriptors may be distinguished:

more general descriptors, which could be the subject of multinational agreement after translation into several languages;

more specific descriptors, which could be the subject of agreement in several special fields.

Descriptors may be:

(a) terms denoting concepts or concept combinations;

(b) terms denoting individual entities. These terms are also called proper names (or identifiers). Proper names may be:

project names
nomenclatures
identification numbers or symbols
geographical or geopolitical names
trademarks
names of persons and organizations
abbreviations and acronyms
other proper names (e.g. programming systems)

It is advisable to use proper names in the same way as other descriptors, i.e. to interrelate them. The same may apply when internationally agreed nomenclatures are integrated into the thesaurus.

In thesauri not using preferred terms, all terms included in the thesaurus may, in principle, be descriptors.

The concept-denoting terms not permitted in indexing must be regarded as unauthorized terms. They are called non-descriptors.

In most cases it would be helpful to provide the possibility of formal distinction between descriptors and non-descriptors. This can be achieved by using a special type for descriptors in print-out or by using special symbols marking the beginning and the end of a descriptor.

3.2 FORMAL REQUIREMENTS

3.2.1 COMPOUND EXPRESSIONS

A descriptor may consist of one or more words. As a general rule, the descriptor should reflect the terminology of the subject, irrespective of the number of separate words required to denote the concept, but it is desirable that the descriptor should contain as few words as possible, and preferably only one. It has to be borne in mind, however, that in the course of abbreviation a term may lose something of its clarity. The words of compound descriptors should be entered in their natural word order (e.g. electrical engineering); i.e. not artificially inverted. It may be helpful to include the inverted forms as non-descriptors preferentially related to the descriptors. When the inverted form is chosen for the entry, i.e. the descriptor, the inclusion of the non-inverted form as a synonym is necessary.

3.2.2 REPRESENTATION OF CONCEPTS BY SEVERAL DESCRIPTORS

To keep the number of descriptors within limits, it may sometimes be useful to represent concepts or combinations of concepts by a combination of descriptors. As a general rule, there are two possibilities:

(a) Morphological factoring: If the concept is represented by a compound or a combination of words it is often possible to break it down into its component parts. When the words thus obtained are combined, however, they only represent the original concept if the morphological factoring coincides with the semantic factoring. Only in this case can a correct result be obtained by morphological factoring.

(b) Semantic factoring: In semantic factoring the descriptors that are to represent the conc t are not necessarily contained in the term.

Example of (a): ANIMAL PSYCHOLOGY = ANIMAL + PSYCHOLOGY

Example of (b): TRADE WIND = METEOROLOGY + GEOGRAPHY

Negative example of a ($\neq$ b): TRADE WIND $\neq$ TRADE + WIND

Terms may be combined (or pre-co-ordinated) before they enter the system and hence must be retrieved as such, or combined during search (post-co-ordinated) to represent a sought concept. Decisions for pre- or post-co-ordination may be taken according to the following considerations:

(a) Post co-ordination

This should be confined to cases where morphological and semantic factoring are identical, the simple descriptors are not used too frequently on their own and/or the precombined descriptor is used very seldom. Care must be taken that the combination of simple descriptors really represents the same concept as the precombined descriptor.

Negative example: TRANSFORMER + OIL = TRANSFORMER OIL
OR OIL TRANSFORMER

(b) Pre-co-ordination

Precombined descriptors should always be used when:

the meanings of the simple descriptors on their own differ from their meaning in the precombined descriptor,
Example: COLLARBONE;

the simple descriptors are used in hierarchical connexions other than the precombined descriptors,
Example: HOUSEWIFE;

the precombined descriptor is a proper name, or doubt exists as to whether the combination of simple descriptors reflects the conceptual content exactly and exclusively.

If a concept is represented by the combination of simple descriptors, this should be expressed by the "Use" reference. The combination of simple descriptors must be included in the systematic sections of the thesaurus, and the unused precombined descriptor in the alphabetical sections.

3.2.3 WORD FORM

Once it has been decided to include a given term in the thesaurus, care should be taken to ensure that, taken together with the relations represented, it conveys the intended meaning as accurately as possible.

(a) Spelling:

The most widely accepted spelling of the word should be used. In cases where, due to varying usage, more than one spelling of a word is accepted, both spellings should be included in the thesaurus and cross-referenced as synonyms.

Example: SULFUR
SULPHUR

Alternatively, a well-established dictionary can be chosen to act as arbitrator whenever this problem arises.

(b) Translation:

Many current technical terms have arisen by translation from other languages, but sometimes a modern foreign language, Latin or Greek term is incorporated into the specialized vocabulary for a particular subject. When both the foreign language term and its putative translation coexist with the same meaning, both should be included in the thesaurus and cross-referenced as synonyms.

Example: BRAKING RADIATION
BREMSSTRAHLUNG

(c) Transliteration:

The problem is further complicated when the foreign language in question is written in a different alphabet. This is particularly true in the case of identifiers. The transliteration standards recommended by the International Organization for Standardization should be used whenever applicable. Wherever a choice exists, the transliteration which does not employ diacritical marks should be selected.

Example: SATELLITE
SPUTNIK

3.2.4 NOUN FORM

Descriptors should be in the form of a noun (or a noun phrase) or that form of the verb which is grammatically equivalent (as the gerund in English).

Example: "DEMOCRACY" instead of "DEMOCRATIC"
"EXECUTION" instead of "EXECUTE"

3.2.5 NUMBER

The use of the singular or the plural form of descriptors should be decided in accordance with the usage in the language of the thesaurus. It is necessary to establish and follow national standards for this decision.

Where no national standard exists, consideration should be given to the following of the lexicographical practice within this language. Sometimes the singular and plural forms of a word denote different concepts, in this case both should be entered.

Example: WOOD
WOODS

In English, in general, the plural form should be used for descriptors, particularly when generic terms are involved (i.e. when the descriptor denotes classes of things).

The singular form is used for specific material or property terms (attributes), process terms, proper names and disciplinary areas.

Examples:

Processes	Properties	Classes of things
ACIDIFICATION	CONDUCTIVITY	TEETH
CALENDERING	OPACITY	STARS
CURING	TEXTURE	PAINTINGS

3.2.6 ADJECTIVES

There are, of course, a certain number of cases where only adjectives or other non-noun forms can be used.

Example: SOCIAL
INTERNATIONAL

A small proportion of single-word terms in adjectival form may be useful as modifiers (continuous, horizontal). Since adjectives can be pre-co-ordinated with nouns and entered as compound descriptors, the choice to enter adjectives singly should be dictated by considerations of practicability and flexibility. Pre-co-ordination is recommended whenever a modifier appears very frequently in combination with another particular term.

3.2.7 ABBREVIATIONS AND ACRONYMS

In general, abbreviated forms of terms should be avoided because their use may not be general enough, their meaning may be dependent on context, or their recognition may be dependent on capitalization and periods which become constraints if computer printers or other electronic data processing equipment is used in conjunction with the thesaurus. Therefore they should be used only when their meanings are well established within the group of users concerned or their meaning is internationally established and when significant gains in practicality can be demonstrated. Abbreviated and unabbreviated forms of a given term should be treated as synonyms and cross-referenced accordingly.

Abbreviations with several meanings are to be treated as homonyms. Sometimes the necessity of limiting the length of the descriptor entails the use of less well established abbreviations. In all these cases a scope note should be appended. Well established acronyms are acceptable as descriptors, e.g. radar, laser.

3.2.8 USE OF CHARACTERS

(a) Character set

The eventual use of electronic data processing equipment may entail:

the use of only the upper-case format for the descriptors

avoidance of diacritical marks

limitation of the number of characters that a descriptor may have

These restrictions will probably disappear in the near future when the electronic data processing equipment is better adapted to the needs of information.

(b) Punctuation

Punctuation marks in descriptors should be minimized. Except for specialized nomenclature, only parentheses and the hyphen are needed in descriptors, as specified in the following guidelines.

Full stops should be allowed only, when due to a limit on the length of the descriptor, a word has to be truncated. Hyphens should be used only when their omission would alter the intended meaning of the descriptor. Commas, colons and apostrophes should be excluded since they are not necessary to convey the meaning of the terms. Where punctuation marks are omitted, it is advisable to include them in full in the scope notes.

Example: LIGHT-SENSITIVE DEVICES
HIGH-VOLTAGE PYLONS

(c) Special characters and numerals

When it is considered necessary to use special characters other than hyphens and parentheses in descriptors, their meanings should be clearly defined. Special characters other than those mentioned above may be used in scope notes, definitions and other forms of additional information, always within the limits machine character availability. If the descriptors contain numerical elements, arabic numerals should be used. The position of the numerals should follow normal usage. Rules must be established for the treatment of subscript and superscript numerals. In the particular case of data retrieval thesauri, the stroke ("/") may sometimes be found necessary.

3.3 METHODS OF AVOIDING AMBIGUITY

3.3.1 HOMONYMS

The different meanings of homonyms (homographs) must be marked and distinguished by specifying symbols or terms (qualifiers) which should be placed between parentheses immediately after the homonym as part of the descriptor. They can be specifying terms, being themselves not homonyms, or other usable signs. Homonym and bracketed qualifiers form a compound descriptor.

Example: BEAMS (ELECTRO-MAGNETIC)
BEAMS (STRUCTURAL)

When a term defined in an internationally or nationally standardized technical vocabulary is selected as a descriptor, it should be written without substantial change, including those parts of the term shown between brackets in the corresponding entry of the vocabulary.

3.3.2 SCOPE NOTES

A scope note is a brief explanation of the intended use of a descriptor. Scope notes may be used:

to restrict the usage of a descriptor

to explain abbreviations and acronyms

to exclude a possible meaning from a term, especially for terms which are in common use in different disciplines

to date addition and deletion of terms and to record changes in the handling of terms.

They may accompany the descriptor in the main part of the thesaurus, but do not form part of the descriptor. Scope notes should be indicated by special characters and clearly distinguished from qualifiers.

Examples: COPPER ALLOYS (alloys in which copper is the principal constituent)
MICROWAVE FREQUENCIES (1 to 300 gHz)
MAMMALS (only the species on land)

3.3.3 DEFINITIONS

The conceptual content of a descriptor in a thesaurus is indicated mainly by represented relations to other thesaurus words. Whenever there is doubt regarding the unique interpretation of a descriptor, a definition should be added, specifying the exact conceptual content.

Example: DOCUMENTATION /the process of storing and retrieving information in all fields of learning/

3.3.4 TRANSLATIONS

In many cases it will be helpful to show the equivalent terms in other languages to ensure that the descriptor is correctly used in the analysis of foreign language texts. Where the meaning is not entirely equivalent, attention can be drawn to this in the form of an explanatory note. Translations should be treated as synonyms or quasi-synonyms.

3.3.5 SOURCE INFORMATION

Information on the source of a descriptor or a definition can be very important for the further development of the thesaurus. The source information should therefore be collected together with the descriptors but need not be included in the printed main part of the thesaurus.

3.4 DESCRIPTOR INTERRELATIONSHIPS

3.4.1 GENERAL

By definition, an indispensable function of a thesaurus is to represent the interrelationships between concepts by representing the interrelationships between the words used to denote them. The network of relations of one descriptor to other descriptors thus provides a kind of definition by placing the descriptor into the semantic space.

Up to now three types of interrelationships require the attention of thesaurus builders: equivalence relation, hierarchical relation, associative relation. All three have the property of reciprocity, i.e. when two or more descriptors are related in any way, reciprocal entries are required.

3.4.2 EQUIVALENCE RELATION (PREFERENTIAL RELATION)

When terms are regarded as equivalent (similar or almost the same in meaning), they can be combined into equivalence categories so that equivalent terms are assigned to one and the same concept. In retrieval, all documents associated with the equivalence category must be retrieved even if only one of the terms is used as the descriptor. A distinction must be made between:

synonyms, i.e. terms which have the same or almost the same meaning in a particular discipline:

change
alteration

and

quasi-synonyms, i.e. terms whose meanings may differ in a specialized vocabulary, but which are considered as synonyms for the purpose of the documentation system under consideration.

(a) USE - reference

In systems using preferred terms, the USE reference is employed to refer from a synonym or quasi-synonym (non-descriptor) to the preferred term (descriptor).

This reference leads the thesaurus user from a term that is not an authorized term to one which is authorized, as follows:

to indicate a preferred synonym: FLEXING USE BENDING

to refer from a specific term to a more general term which has been selected to represent the specific concept (quasi-synonym): PLANT WAXES USE WAXES

to indicate a preference in spelling or to expand or explain abbreviations: PI MESONS USE PIONS

to prescribe the use of two or more terms to express a concept (semantic factoring): FERROMAGNETIC FILMS USE FERROMAGNETIC MATERIALS + FILMS

to express concepts that can be considered synonyms for purposes of indexing and retrieval (quasi-synonyms): HEREDITY USE GENETICS

to bring together different points or degrees of a conceptual continuum: FLUIDITY USE VISCOSITY

to cross-reference indirect entries to the preferred natural word order: TABLES, mathematics USE MATHEMATICAL TABLES

to reflect current terminology: ELECTRICAL CONDENSERS USE CAPACITORS

to provide preferred terms for jargon: WHIRLY BIRD USE HELICOPTERS

to include translations of descriptors: ROENTGENSTRAHLEN USE X-RAYS

(b) USED FOR - reference

Conversely, the USED FOR reference is employed with the preferred term for reciprocal reference. It accompanies the term to which the USE-reference refers.

3.4.3 HIERARCHICAL RELATION

The hierarchical relation expresses relations of super-/subordination of concepts. It can be subdivided into:

(a) The generic relation
in which the generic (superordinated) term denotes a class of concepts of which the concept denoted by the specific (subordinated) term is a member. The specific concept differs from the generic one in at least one characteristic.

Example:

INTERNAL COMBUSTION ENGINE

RECIPROCATING-PISTON ENGINE	OTTO CYCLE ENGINE
ROTARY-PISTON ENGINE	DIESEL ENGINE

(b) The part-whole-relation (partitive relation)
in which the superordinated term (entity term) denotes objects or concepts of which the objects or concepts denoted by the subordinated term (part term) are a part. Part terms are obtained by mental disintegration of the whole represented by the entity term into its component parts.
Example:

Entity term: RECIPROCATING-PISTON ENGINE

Part terms:
CYLINDER
PISTON
CONNECTING ROD
CRANKSHAFT
etc.

Hierarchical relations can be represented in any of the following ways:

The generic and part-whole-relations are differentiated and shown separately.

The generic and part-whole-relations are not differentiated and are grouped together in the hierarchical reference.

In disciplines in which the part-whole-relations are of no significance in hierarchical retrieval, it is recommended that only the generic relation be represented by hierarchical reference. In this case the part-whole-relation is treated as associative relation.

The representation of the part-whole-relation by both hierarchical and associative reference in a thesaurus is to be avoided.

In most thesauri the hierarchical relation is represented by the reference BROADER TERM (BT), representing the relation of a concept being superordinated, and NARROWER TERM (NT), indicating the reciprocal relation. In cases where both types of hierarchical relations are to be distinguished, different symbols must be chosen for generic and part-whole-relations. It is recommended to use the following references:

BROADER TERM GENERIC (BTG) and NARROWER TERM GENERIC (NTG) for the generic relation, and BROADER TERM PARTITIVE (BTP) and NARROWER TERM PARTITIVE (NTP) for the part-whole-relation.

3.4.4 ASSOCIATIVE RELATION (AFINITIVE RELATION)

The associative relation usually employed to cover the other relations between concepts that are closely related but are neither consistently hierarchical nor equivalent (e.g. similarity, antonymity). It should be noted, however, that a variety of relations exist between concepts. Associative relations should therefore be established only if it is assumed that these relations will be actually required in retrieval. Associated concepts can be referred to by the RELATED TERM (RT) reference. Associative relations may be used to indicate, for example:

antonymity, i.e. a concept is the opposite of another concept:
HARDNESS RT SOFTNESS

co-ordination, i.e. concepts are derived from a superordinated concept by the same step of division:
GENERIC RELATION RT PART-WHOLE-RELATION

genetic relation, i.e. something is the predecessor of another thing:
FATHER RT SON

concurrent use of two concepts: EDUCATION RT TEACHING

cause and effect: TEACHING RT LEARNING

instrumental relation:
WRITING RT PENCILS

material relation, i.e. something is the material of which another thing is made:
PAPER RT BOOKS

similarity of different kinds (physical similarity, similarity of material, similarity of processes, etc.):
TEACHING RT TRAINING

This list is intended only to illustrate the various kinds of relations which may be covered by the associative relation. Only those relations between two terms that prove to be of sufficient effect in indexing or retrieval should be established.

In particular fields it may be useful to represent some of these relations explicitly, besides the hierarchical and associative relation. These additional relations should be clearly defined and coded.

3.4.5 SYMBOLS FOR CROSS-REFERENCES

The interrelationships can be expressed by several means. If codes are used to indicate these relationships, their meanings should always be made clear. It is recommended to use the following abbreviations:

BT BROADER TERM (when generic and part-whole-relations are not to be distinguished.)

NT NARROWER TERM (when generic and part-whole-relations are not to be distinguished)

BTG BROADER TERM GENERIC (generic term)

NTG NARROWER TERM GENERIC (specific term)

BTP BROADER TERM PARTITIVE (entity term)

NTP NARROWER TERM PARTITIVE (part term)

RT RELATED TERM (associative relation)

USE USE synonym, quasi-synonym, or a combination of descriptors

UF USED FOR a synonym, quasi-synonym,, or a combination of descriptors

When this recommendation cannot be adopted by a country, this country should use only one type of abbreviations compatible with the above recommended abbreviations (according to rules adopted in the country).

4. PRESENTATION AND ARRANGEMENT OF THE THESAURUS

4.1 INTRODUCTION TO THE THESAURUS

No thesaurus should be presented without a comprehensive introduction which states clearly the purpose and structure of the thesaurus, and the domains covered by it. The rules followed in its establishment should be presented in a condensed form. This is particularly true of the methods and sources used in the selection, form and avoidance of ambiguity of the descriptors. The method of presenting the thesaurus, the meaning of the abbreviations used and the rules for alphabetization and punctuation, whenever applicable, should be explicitly stated. Most important of all, the rules for using the thesaurus and its limits of applicability should be elucidated and illustrated by means of examples, where appropriate.

Users should be invited to contribute comments and suggestions for the improvement of the thesaurus, and to inscribe themselves on the mailing list for future editions of or additions to the thesaurus. The proposed system for developing and updating the thesaurus should be explained; the date of the present, and estimated appearance of future editions or additions to the thesaurus should be given.

The total number of descriptors and non-descriptors should be stated.

4.2 MAIN PART OF THE THESAURUS

A thesaurus should always include a systematic and an alphabetical display. The main part should comprise complete information on each descriptor (in systems using preferred terms) or concept (in systems not using preferred terms). The main part can be arranged systematically or alphabetically, or a combination of both can be used. When non-descriptors and permutations of compound descriptors and non-descriptors are included in an alphabetically arranged main part, a separate alphabetic index is not required. The information to be included in the main part comprises:

Concept representations

Concept number and/or notation(s)
Descriptor(s)
USED FOR - references (equivalence relations)

Additional information

Definition(s)
Scope notes
Short source information

Concept relationships

Superordinated concepts (hierarchical relations) - BT
  Broader concepts (generic relation) - BTG
  Entity concepts (part-whole relation) - BTP

Subordinated concepts (hierarchical relations) - NT
  Narrower concepts (generic relation) - NTG
  Part concepts (part-whole relation) - NTP

Other specified relationships

Related concepts (associative relations) - RT

It should be noted, however, that thesauri are polyhierarchical and/or facet-structured so that there may be several broader or narrower concepts for a given term. Since only one hierarchical chain can be presented coherently in a systematic display, the relationship with other hierarchical chains must be indicated by cross-reference.

4.3 AUXILIARY PARTS OF THE THESAURUS

4.3.1 GENERAL

In order to improve access to the main part of the thesaurus, a thesaurus may contain several auxiliary parts, i.e. alphabetical indexes, systematic listings and graphic displays. An alphabetical index is required when:

the main part is arranged systematically

the main part is arranged in a combination of systematic and alphabetical listings

the main part arranges only the descriptors (in systems using preferred terms)

A systematic listing is required when:

the main part is arranged alphabetically

the main part is arranged in a combination of systematic and alphabetical listings.

The auxiliary parts refer to the appropriate entry in the main part. The use of consecutive descriptor or concept numbers or of a notation is useful for this purpose.

4.3.2 ALPHABETICAL INDEXES

Alphabetical indexes should comprise descriptors and non-descriptors. Permutation of compound descriptors and compound expressions is desirable. The alphabetical index can be arranged in the form of a permutation or keyword-in-context.

The national and international standards should preferably be used for alphabetization. Since the various methods of alphabetization are used differently in different domains, the rules used for alphabetic arrangement must be stated explicitly before sorting is started.

4.3.3 SYSTEMATIC LISTINGS

Systematic listings should contain all descriptors (in systems using preferred terms) or all concepts (in systems not using preferred terms) in accordance with the hierarchical relationships represented in the thesaurus. If more than one type of hierarchical relationship is used in the thesaurus, the relations can be combined in one listing or indicated in separate listings. Systematic listings should show the various hierarchical levels by appropriate indentations. As the thesaurus is always polyhierarchical, it is necessary to list terms which are common to several hierarchical chains several times at the appropriate places in the monohierarchical representation. If a term is hierarchically related to more than one broader term on the next higher level, it is advisable to list the other relevant broader terms as well, marking the hierarchical level by an appropriate symbol (e.g. +) and an indentation to the right.

Example:

METAL WORKING

COLD WORKING

METAL ROLLING

COLD ROLLING

ROLL PLANISHING

SURFACE ROLLING

This structure can be represented in a systematic index as follows ("-" = subordination, "+" = superordination):

METAL WORKING

- COLD WORKING

- - COLD ROLLING

+ METAL

+ + METAL WORKING

- - - ROLL PLANISHING

- - - SURFACE ROLLING

- METAL ROLLING

- - COLD ROLLING

+ COLD WORKING

+ + METAL WORKING

- - - ROLL PLANISHING

- - - SURFACE ROLLING

This display method permits the polyhierarchical structures to be restored. Systematic listing is probably better for very specialized scientific and technical fields than for interdisciplinary areas.

A combination of this type of display with alphabetical listing gives rise to a kind of structured alphabetical list which probably combines to the fullest extent the advantages of both.

4.3.4 GRAPHIC DISPLAYS

Perhaps the most subtle mode of presentation of thesauri is to display the descriptors and the relationships between them graphically. Although this can be done multi-dimensionally, for instance by taking two dimensions for each facet of a multi-faceted thesaurus, the more current methods are two-dimensional. One such system consists of arranging the descriptors in semantic groups, assigning a gridded sheet to each group and giving fixed positions to each descriptor with respect to the horizontal and vertical axes, thus defining co-ordinates.

Interrelationships between descriptors in graphic displays can be shown by:

graphic arrangement (formation of groups, entry in a system of co-ordinates, use of different abstraction levels, symbolic concentric circles indicating different hierarchical levels);

arrows between descriptors (bi-directional arrows for associative relationships, uni-directional arrows for hierarchical relationships pointing to the more specific descriptor, brackets with the arrows leaving or arriving at the preferred term for equivalence relations);

It is understood that a descriptor may belong to several groups. The optimal size of each group appears to lie between 30 and 40. As before, an alphabetical listing should be given in the annex showing the semantic group(s) on which each descriptor belongs. Which mode of presentation is selected will depend on the use to which the particular thesaurus will be put.

4.4 COLLATING SEQUENCE (ALPHABETIZATION)

For alphabetic listings in the main part of the thesaurus or in the alphabetical index a collating order for the order of letters, numbers, punctuation marks and special symbols must be defined. The arrangements may be:

(a) Letter-by-letter. A typical collating sequence is based on the following rules:

Ignore all spaces between words.

Ignore all characters other than left parentheses, numerals and letters.

File according to the sequence: left parenthesis, numerals in ascending value, letters in order A-Z.

(b) Word-by-word. A typical collating sequence is based on the following rules:

Each complete word is considered in turn and arranged in letter-by-letter sequence.

All terms beginning with a given complete word precede any terms beginning with the same sequence of letters as part of the word.

Non-alphanumeric characters (punctuation marks and special characters) are treated as spaces.

(c) Computer sort. A typical collating sequence would be based on the following rules:

All characters, including letters, numerals and punctuation marks and special characters within a term are considered in determining the sequence.

The number of spaces within a term must be fixed.

Numerical sequence, not numerical value, is followed.

The exact sequence would be based on the specific computer configuration used. The selection of the method of alphabetization depends on all the factors affecting the thesaurus under construction, i.e. the size and structure of the domains covered by the thesaurus, the availability of machine processing the kind of hardware available, etc. In all cases the alphabetization rules should be clearly and explicitly drawn up before any kind of ordering is attempted.

5. THE PROCESS OF THESAURUS CONSTRUCTION

5.1 AVOIDANCE OF DUPLICATIVE WORK

Before commencing work on the establishment of a thesaurus, it is advisable to ascertain whether others covering that particular domain, or a neighbouring one, are available. This is best done by addressing a query to one of the two clearinghouses.(*) It may be found advisable to go ahead with the compilation of a particular thesaurus in spite of the usage of a similar one. In this case, the reasons for proceeding, as well as the differences between the new and the earlier thesaurus should be stated clearly in the introduction.

5.2 NOTIFICATION OF INTENT

The appropriate clearinghouse (see above) should be notified of the intention to construct a thesaurus, as well as when the thesaurus is first published or disseminated. This information should be channelled through the national organization dealing with thesauri, where and when such an entity exists.

The same applies for further editions. If at all possible, a copy of the thesaurus, complete with the introduction and indexes should be sent to the clearinghouse in question. The fact of notification should be mentioned in the introduction.

5.3 GENERAL PRINCIPLES

Before the general structure of a thesaurus is agreed upon a survey should be made of the structure of the specialized terminology involved and of the needs of the users of the information and documentation system. The selection of descriptors should start only after the general structure of the thesaurus has been agreed upon. It is recommended:

to entrust the more time-consuming parts of the work and the co-ordination of the work done by subject specialists to a few full-time workers;

to include as many experts as possible having both a good knowledge of the subject field to be treated and previous experience in indexing and retrieval.

to form panels working in special areas of the subject field to be covered.

to use internationally recruited teams, since this widens the cumulative linguistic experience which goes into the building of the thesaurus.

(*) for English-language thesauri:
Bibliographic Systems Center, School for Library Science, Case Western Reserve University, Cleveland, Ohio 44106, U.S.A.

for thesauri compiled in languages other than English:
Centralny Institut Informacji Naukowo Technicznej i Ekonomicznej (CIINTE) Clearinghouse Al. Niepodległości 188, Warzsaw, Poland.

The methods of selecting descriptors vary according to the proposed structure of the thesaurus (alphabetical listing, systematical listing, graphic display), the purposes for which the thesaurus will be used (e.g. for manual or mechanical retrieval, only for indexing, or as a secondary tool) and the background to the project (gradual build-up to mechnical processing, introduction of a new domain, e.g. interdisciplinary areas for which no previous classification schemes existed; the existence of a well-defined group of users and subject specialists; extensive literature).

5.4 DIFFERENT APPROACHES

In general, there are two approaches to thesaurus construction. The choice between them is made on the basis of the type and scope of the literature or the collection which the thesaurus is intended to cover. These two approaches may be called the analytical method and the gestalt method. Both are empirical in that neither depends on rigorous linguistic analysis. In general, the analytical method involving analysis of the subject content of the literature and the selection of terms from the literature, is to be preferred, especially for specialized areas of knowledge. The gestalt method employs experts who analyse candidate terminology from secondary sources such as indexes and other reference materials and who make selections of preferred terms and determine the relationships of terms to each other. The gestalt method is more generally applicable to broad subject fields involving several disciplines. In actual practice, a combination of both approaches is ofter used. The analytical approach starts with indexing a collection of documents comprising the field of knowledge the thesaurus is intended to cover. Using the data obtained by this process of experimental indexing, the thesaurus builder can start to establish tentative relationships valid for the retrieval system concerned. Then the thesaurus will be corrected by reindexing the starting collection and expanding the material to be indexed.

The gestalt approach needs as a starting point an overview of the structure of terminology in the area of knowledge concerned, the structure being in effect, a taxonomy or classification. Experts familiar with the subject matter are asked to recognize such relationships and structures as they review terms experimentally selected for describing documents. In contrast to the analytical approach, the thesaurus is not built up by practical experience, but by the contributions of "authorities".

5.5 SELECTION OF DESCRIPTORS

In general, four distinct steps intervene in the selection of descriptors: collection, verification, evaluation and choice.

5.5.1 COLLECTION

It is almost impossible to make a comprehensive collection of candidate terms by thinking of an alphabetical list. By envisaging descriptors in groups,

thought associations between them give rise to many candidates. Sources for term collection(*) may be:

potential users and subject specialists

internationally or nationally standardized dictionaries

current literature

terminological treatises or lists

existing thesauri and classification schemes

nomenclatures

indexes of journals

abstracting services

textbooks, handbooks and summaries

experimental indexing of documents

Certain fields have highly specific systems of nomenclature, or well-established standardized technical vocabularies. Whenever an internationally agreed nomenclature exists, it should be used. The proliferation of unrelated specific names would tend to convert the thesaurus into a simple list of identifiers which would be self-defeating. It is therefore recommended that the names of unrelated specific entities be avoided as much as possible. Specific names which are not treated as descriptors (i.e. by showing their interrelationships) may be listed separately from the main part of the thesaurus.

5.5.2 VERIFICATION

With all methods of assembly, the authenticity of the selected descriptors should be verified by consulting dictionaries, other indexing or standardized vocabularies, current usage in the literature and especially the opinion of subject specialists. Obsolete terminology should not be included, if so only as forbidden terms.

One of the more appealing attributes of a thesaurus is its ability to assimilate immediately the neologisms and special jargon that proliferate in expanding fields of basic and applied research. Full advantage should be taken of this facility in combination with the use of scope notes and cross-references. Special care should be taken with terms whose connotations have changed with the passage of time, or whose meaning changes from country to country. If overlapping terms must be included, the appropriate cross-reference should be employed.

(*) Reference may be made to the International Information Centre for Terminology, located at the Osterreichisches Normungsinstitut (Austrian Standards Institute) Vienna, Austria.

5.5.3 EVALUATION

In evaluating the utility of candidate descriptors, reference should be made to their: (1) frequency as encountered in the literature or in the existing stocks of information; (2) anticipated incidence in retrieval inquiries; (3) relationship to descriptors already accepted; (4) appropriateness and authenticity as current terminology in the discipline concerned; (5) effectiveness and expediency in connoting and denoting the particular concept. None of these factors should be considered independently and particular attention should be paid to areas of peripheral interest where the exhaustivity and specificity required of the descriptors are not the same as for the core subject.

5.5.4 CHOICE

In all cases, descriptors should be selected for the inclusion in the thesaurus on the basis of their estimated effectiveness for retrieval purposes and their measureable significance in the material to be indexed.

5.6 RECORDING PROCEDURES

The procedures for recording information during the construction process may vary according to the specific requirements of each system. (see 5.3.). Therefore only some practical indications can be given.

The process of selection of terms and establishment of relations between them makes it necessary to record the information collected on a specific term in a formalized manner. Therefore, in this process a format must be developed and used specifying those items of information which are used for thesaurus construction (see 4.2.). The physical appearance of the format depends on the system intended and the equipment available. It may be a form, a manually operated card, a punched card, or a set of categories with variable length usable for computer input. Each time a term is considered to be a candidate descriptor or non-descriptor a record must be established. The concept relations contained in the sources from which the candidate term has been selected should be recorded during this selection process. In each case it is necessary to determine whether the relations found within the source depend on the specific context or hold in general.

During construction, tentative decisions can be made and changed later; this provides desirable flexibility, but it requires the builder to make proper notations under every term affected by each decision. For example, the construction "Term A USE Term B" requires the USE note on the term A record and the reciprocal USED FOR note of the term B record. The same applies to the other kinds of references. Following this procedure, it is possible to change a decision at any time; the note structure provides the information necessary to keep the thesaurus internally consistent; all affected entries can be located and changed to conform to a new decision.

In deciding on the entry of descriptors, the selected terms should be grouped systematically, e.g. according to facets. The concepts denoted by the descriptors should be clarified. Each descriptor should be checked for its concept relationships with other descriptors (in accordance with the proposed thesaurus structure). Subject specialists and potential users should be consulted in selecting the descriptors and determining the concept relationships.

5.7 PILOT RUN

Before establishing a thesaurus on a definitive basis, it is strongly recommended that a practical test based on a sufficient number of documents dealing with the whole area subsequently to be covered be carried out. This pilot run should show:

whether the proposed thesaurus structure meets the later requirements,

whether the selected descriptors are useful for indexing and retrieval,

in which areas further descriptors are required, and

whether the established concept relationships are correct and sufficient.

The results of this test should be critically commented upon by the greatest feasible number of people, including information scientists and indexers as well as subject specialists and users.

6. UPDATING AND MAINTENANCE

The thesaurus should be updated continuously on the basis of the questions raised in 5.7. For this, a central authority should be nominated and charged with the updating of the thesaurus according to a defined procedure.

6.1 PERIODIC VERIFICATION

Periodically a check should be kept on the frequency with which particular descriptors are utilized, both for indexing and retrieval purposes. Periodic verification should ensure that certain descriptors neither interfere with, nor duplicate one another, and that the relations established within the thesaurus are still right and practicable. On all occasions in which a search does not locate the desired information or the amount of information suspected of being in the collection, a critical appraisal of the descriptors which were, or should have been used, ought to be carried out.

Descriptors which, due to changes in the subject field and terminology, have become obsolete, should be eliminated or replaced by more appropriate descriptors.

6.2 ELIMINATION OF DESCRIPTORS

If it is found that a descriptor is used very infrequently, it should be ascertained whether the infrequent usage is due purely to the lack of documents related to the particular concept. Complete elimination, in principle, should occur only when that particular descriptor has never been used, either for indexing or retrieval purposes. The use of a preferential relationship (USE reference) to indicate where the replacement has been effected is more practical.

If a descriptor has been eliminated, it may be retained as a synonym or be allowed for retrieval only to avoid the necessity of reindexing documents in which the eliminated descriptors have been used. In each case the relationships of the deleted or changed descriptor must be examined and adapted to the new situation.

If, on the other hand, too many indexed materials are assigned to the same descriptor, its specificity is lost, its application has become too general and the breaking-down of the concept should be considered. The procedure to be followed when a particular descriptor is over or underused depends to a certain extent on the search strategy employed in retrieval. If the least specific descriptor is searched for last, it may not be worthwhile to eliminate it.

6.3 CHOICE OF NEW DESCRIPTORS

If, during indexing and/or retrieval, it is found that concepts or concept relationships have not been established with sufficient precision in the thesaurus, new descriptors or relationships must be established. Furthermore, indexers and users should constantly be on the look-out for new candidate descriptors which may represent either new concepts or different facets of old concepts or which are used as synonyms to existing terms. New terms and relations must be evaluated for usefulness before they are definitely entered.

If possible, the new descriptor should be used by indexers on a trial basis for some time before it is definitely added to the thesaurus.

The frequency of occurrence of such candidate descriptors both as indexing and retrieval terms is a good indication of their future usefulness. If it is decided to add a new descriptor, the interrelationships with all the pre-existing descriptors should be identified and introduced in the appropriate places.

Definite additions should not be introduced singly as this causes confusion among the users of the thesaurus. New descriptors should be saved up and introduced by batches, either as "additions to the thesaurus" or on the occasion of a new edition of the thesaurus. This does not preclude their use by indexers. There should exist a central authority which examines all the suggestions received and issued a final verdict on the acceptability or otherwise of the possible new additions.

It should always be remembered that a thesaurus is never completed, its size and shape being a function of time.

Any comments on the above text may be sent to the Department of Science Policy, Division of Scientific Documentation and Information, Unesco, Place de Fontenoy, Paris VII, France.

LIST OF ISO RECOMMENDATIONS RELATED TO THESE GUIDELINES

ISO/R 9 - 1968	International system for the transliteration of Slavic characters.
ISO/R 18 - 1955	Short contents List of Periodicals or other documents.
ISO/R 30 - 1956	Bibliographical strip.
ISO/R 77 - 1958	Bibliographical references. Essential elements.
ISO/R 214 - 1961	Abstracts and Synopsis.
ISO/R 233 - 1961	International system for the transliteration of Arabic characters.
ISO/R 259 - 1962	Transliteration of Hebrew.
ISO/R 639 - 1967	Symbols for languages, countries and authorities.
ISO/R 646 - 1967	6 and 7 bit coded character sets for information processing interchange.
ISO/R 690 - 1968	Bibliographical references. Essential and supplementary elements.
ISO/R 704 - 1968	Naming principles.
ISO/R 832 - 1968	Abbreviations of typical words in bibliographical references.
ISO/R 833 - 1968	Abbreviations of generic names in titles of periodicals.
ISO/R 843 - 1968	International system for the transliteration of Greek characters into Latin characters.
ISO/R 860 - 1968	International unification of concepts and terms.
ISO/R 919 - 1969	Guide for the preparation of classified vocabularies.
ISO/R 999 - 1969	Index of a publication.
ISO/R 1087 - 1969	Vocabulary of terminology.
ISO/R 1149 - 1969	Layout of multilingual classified vocabularies.
ISO/R 1682 - 1971	Dimensions and location of rectangular punched holes in 80 columns punched paper cards.
ISO/R 1861 - 1971	7-track 8 RPmm (200 RPI) magnetic tape for information interchange.
ISO/R 1862 - 1971	9-track 8 RPmm (200 RPI) magnetic tape for information interchange.
ISO/DR 1951 (Draft)	"Lexicographical symbols, particularly for use in classified defining vocabularies".

The above documents are available either from the Headquarters of ISO (International Organization for Standardization), 1 Rue de Varembé, Geneva 20, Switzerland.

or from: the corresponding National Standards Organizations for the member countries of ISO.

Sources for dictionaries, glossaries and thesauri

Bibliography of interlingual scientific and technical dictionaries, 5 ed. Paris, Unesco, 1969. 250 p.

Bibliography of monolingual scientific and technical glossaries, Vol. I: National Standards, 1955, 210 p. Vol. II: Miscellaneous Sources, 1959, 146 p. Paris, Unesco.
(Supplements published in Babel, International Journal of Translation published by the International Federation of Translators with the assistance of Unesco. Avignon, France).

International Information Centre for Terminology, affiliated to the Austrian Standards Institute, Vienna, Austria.

Bibliographic Bulletin of the Clearinghouse at CIINTE, 1969, Warsaw, CIINTE, 1969, 140 p. (with annual supplements)

Bibliographic Systems Center Subject Index, Case Western Reserve University, Cleveland, Ohio, U.S.A., 1969. (Computer print-out).

Bulletin des Bibliothèques de France, 14ème année, no. 5, mai 1969, pp. 203-218
" " " " , 15ème année, no. 1, janvier 1970, pp. 21-26
" " " " , 16ème année, no. 1, janvier 1971, pp. 33-38

Kathleen Gaster
Thesaurus construction and use of selective bibliography based on material in the Aslib Library in July 1967, Aslib Proc. Vol. 19/1967/310-317.

J.F. Blagden
Thesaurus compilation methods: A literature review. Aslib Proc. Vol. 20/1968/ S.345-359.

I.L. Barske, L. Tschache:
Thesauren und ähnliche Begriffslisten. Berichtszeit: 1960-1968. Dresden: TU Dresden 1969 (Bibliothek d. TU Dresden, Bibl. Arb. Nr. 4).